光电技术及应用

DOI：10.16818/j.issn1001－5868.2022042701

基于多智能体深度强化学习的智能电网光网络切片方案

戚银城[1*]，唐奕明[2]

（华北电力大学 1. 电气与电子工程学院；2. 国际教育学院，河北 保定 071003）

摘　要： 为了提高光网络对大规模、差异化电力业务的资源分配能力，降低大规模业务的算法训练时间，提出了一种基于多智能体深度确定性策略梯度算法的智能电网光网络资源分配方案。该方案考虑大规模和差异化电力业务，将智能电网光网络建模成多智能体系统，以最大化电网公司收益为目标，建立了智能电网光核心网络切片模型，进行网络资源分配优化，并采用条件判断映射，简化了优化问题。同时，把不同业务部署到不同智能体中进行运算，以降低训练时间，满足网络实时性需求。仿真结果表明，该算法具有更大的奖励、更低的成本、时延和训练时间。

关键词： 智能电网；光网络切片；资源分配；多智能体；条件判断映射

中图分类号： TN929.14；TM7　**文章编号：** 1001－5868(2022)05－0979－07

Smart Grid Optical Network Slicing Scheme Based on Multi-agent Deep Reinforcement Learning

QI Yincheng[1]，TANG Yiming[2]

（1. School of Electrical and Electronic Engineering；2. School of International Education，North China Electric Power University，Baoding 071003，CHN）

Abstract： In order to improve the resource allocation ability of optical networks for massive and differentiated power services and reduce the algorithm training time of large-scale services，a smart grid optical network resource allocation scheme based on the multi-agent deep deterministic policy gradient（MADDPG）algorithm was proposed. The large-scale and differentiated power services were considered，the optical core network slice model of smart grid was built and the optimization problem aiming at maximizing the income of power grid companies was proposed. Conditional judgment mapping was proposed to simplify the optimization problem. At the same time，the improved MADDPG algorithm was designed to reduce the training time and meet the real-time needs of the network by placing different services to different agents. Lastly，simulation results show that the proposed algorithm has better reward，lower cost and time delay，and lesser training time.

Key words： smart grid；optical network slicing；resource allocation；multi-agent；conditional judgment mapping

0　引言

近年来，随着智能电网中光网络通信概念的提出和工业物联网的普及，电力设备入网数量快速增长[1]，大规模和差异化电力业务已成为主流趋势。网络切片作为一种按需组网的方式，由于具有根据业务需求，灵活地对无线网侧协议栈功能进行定制切分的优势，广泛应用于第五代移动通信（5th Generation，5G）网络中[2]。网络切片问题难点在于NP-Hard和非凸优化问题，无法通过常规方法求解。近年来，深度强化学习（Deep Reinforcement Learning，DRL）作为上述解决难点的一种有效方案，在网络切片优化中广泛使用[3]。然而，现有大部分强化学习为集中式设计，受限于智能体的计算能

收稿日期：2022－04－27.

*通信作者：戚银城　E-mail：qiych@ncepu.edu.cn

力和训练时间，不适用于大规模和差异化程度极高的智能电网光网络资源分配。另一方面，由于网络切片需要同时设计虚拟网络功能(Virtual Network Function，VNF)部署和资源分配方案，优化问题时包含离散和连续变量，传统 DRL 无法使用。

通常，智能电网光网络的 VNF 部署和资源分配是一个包含节点、链路以及其中计算、缓存和带宽等资源的高度复杂的综合优化问题，传统方法难以解决。到目前为止，智能电网光网络的 VNF 部署和资源分配还没有完善的解决方案。因此，设计一个考虑多种因素的资源分配策略非常重要，例如系统可扩展性、物联网设备的异构性和互操作性以及资源供应商和客户之间的交易。DRL 的最新进展显示出解决光网络的 VNF 部署和资源分配问题的巨大潜力。作为人工智能(Artificial Intelligence，AI)的一个子组，DRL 可以被认为是强化学习(Reinforcement Learning，RL)和深度学习(Deep Learning，DL)的集成。一方面，RL 涉及专注于最大化长期性能的自学习代理，它从与环境的交互中学习，不需要了解系统模型。另一方面，DL 在高维感官表征的挑战上取得了显著进展。受生物神经网络的启发，DL 构建在通常具有多层的人工神经网络之上。从理论上讲，具有一个或多个隐藏层的多层神经网络有可能成为通用逼近器。因此，通过利用 DL，DRL 能够有效地提取特征并避免维度灾难。

很多文献对网络切片进行了深入研究，并取得进展。文献[4]讨论了 DRL 在无线资源切片和核心网络切片等场景中资源管理的典型应用。文献[5]提出启发式算法和离线强化学习相结合的切片方案。文献[6-8]利用深度 Q 网络(Deep Q Network，DQN)进行动态资源分配。其中文献[6]对带宽比例进行简单分配，文献[7]对资源块进行细化分配，文献[8]提出了改进的 DQN 算法，通过引入离散归一化优势函数，使算法更快收敛。文献[9]提出确保可靠性的最小运营支出的服务功能链(Service Function Chain，SFC)部署方案。文献[10]提出了可靠性感知和时延受限的 VNF 放置和流量控制策略。文献[11]探讨了双连通性(DC)对异构航空网络中提供统计服务质量(QoS)的延迟感知资源控制的影响。文献[12]考虑 5G 场景三种典型切片，采用复杂网络理论对节点进行映射，并选择节点间连接路径。文献[13]构建了面向移动边缘计算的端到端切片方案。文献[14]提出考虑服务质量的端到端网络切片框架。但是，在解决大规模和差异化程度极高的智能电网资源分配时，由于很长的训练时间和复合变量环境，上述文献方案不能使用。

为了解决上述问题，本文提出一种基于多智能体深度确定性策略梯度(Multi-Agent Deep Deterministic Policy Gradient，MADDPG)的智能电网光网络资源分配方案，把智能电网光网络资源分配问题建模成一个多智能体系统。该方案考虑大规模和差异化电力业务，建立智能电网核心网络切片模型，提出最大化电网公司收益为目标的优化问题，并通过优化目标函数来实现网络资源分配。另外，本文还利用条件判断映射，实现了优化问题的简化。同时，本文通过部署不同业务到不同智能体中进行运算，以降低训练时间，满足网络实时性需求。

1 系统模型

1.1 网络模型

考虑一个具有高带宽、低时延和普通业务的光核心网络，网络场景如图 1 所示。该网络抽象为无向图 $\boldsymbol{G}(\boldsymbol{N},\boldsymbol{L})$。其中，$\boldsymbol{N}=\{n_1,n_2,\cdots,n_N\}$ 和 $\boldsymbol{L}=\{l_1,l_2,\cdots,l_L\}$ 分别表示核心网的节点和链路集合。节点集合 $\boldsymbol{N}$ 可以划分为源节点 $\boldsymbol{N}^{\mathrm{S}}$、中间节点 $\boldsymbol{N}^{\mathrm{R}}$ 和目标节点 $\boldsymbol{N}^{\mathrm{D}}$。节点资源由计算资源、存储资源和节点具有的 VNF 集合共同构成。节点 $n\in\boldsymbol{N}$ 的最大计算资源和存储资源分别表示为 C_n 和 B_n。节点 $n\in\boldsymbol{N}$ 的 VNF 集合表示为 $\boldsymbol{O}_n=\{o_n^1,o_n^2,\cdots,o_n^F\}$，其中 o_n^i 表示第 i 种 VNF。链路具有带宽资源。链路 $l=(n,n')\in\boldsymbol{L}$ 的带宽资源表示为 W_l。本文最短时间间隔设置为时隙 τ，并且时隙是业务请求最小变化单位。业务请求集 $\boldsymbol{X}=\{X_1,X_2,\cdots,X_n\}$，其中 $X_s\in\boldsymbol{X}$ 是 SFC，由需要经过的 VNF 序列组成，描述为 $X_s=\{o_s^1,o_s^2,\cdots,o_s^{F_s}\}$。其中 o_s^i 是业务 s 所经过的第 i 个 VNF。业务 s 的第 i 个 VNF o_k^i 所需计算、内存和第 l 条链路带宽资源分别表示为 $c_{s,\mathrm{need}}^i=R_s c_i$，$b_{s,\mathrm{need}}^i=R_s b_i$ 和 $w_{s,\mathrm{need}}^l=R_s/\eta_l$。其中 R_s 表示业务 s 的比特速率，c_i 和 b_i 分别表示 VNF 执行 1 bit 计算或存储使用的 CPU 周期和内存，η_l 表示链路 l 的频谱效率。

1.2 VNF 部署

对于 VNF 部署，主要包括 SFC 中 VNF 和链路的确定以及计算功能的部署。对于前者，使用 $u_{i,i'}^{\mathrm{s}}$ 表示 VNFi 和 i'，以及链路 $l=(n_i,n'_i)$ 需要转发和

传输流量；对于后者，使用 v_i^s 表示需要执行计算和存储功能。此外，需要通过以下约束限制 VNF 部署和链路选择。

$$\sum_{i\notin \boldsymbol{N}_s}\sum_{n'\in \boldsymbol{N}} u_{i,i'}^s(t)=0 \tag{1}$$

$$\sum_{n\notin \boldsymbol{N}_i} v_i^s(t)=0 \tag{2}$$

$$\sum_{n\in \boldsymbol{N}_i}\sum_{n'\in \boldsymbol{N}} u_{i,i'}^s(t)=1 \tag{3}$$

$$\sum_{n\in \boldsymbol{N}_i} v_i^s(t)\leqslant 1 \tag{4}$$

$$\sum_{n'\in \boldsymbol{N}} u_{i,i'}^s(t)\geqslant v_i^s(t) \tag{5}$$

其中，$\boldsymbol{N}_s$ 表示包含业务 s 的 SFC 所需 VNF 的节点，式(1)～(4)分别表示业务 s 的流量转发 VNF 和功能执行 VNF 必须在业务 s 的 SFC 所需 VNF 的节点中执行，并且一个节点只能使用一个 VNF 执行传输或计算过程。式(5)表示执行计算和存储的 VNF 必须转发流量。

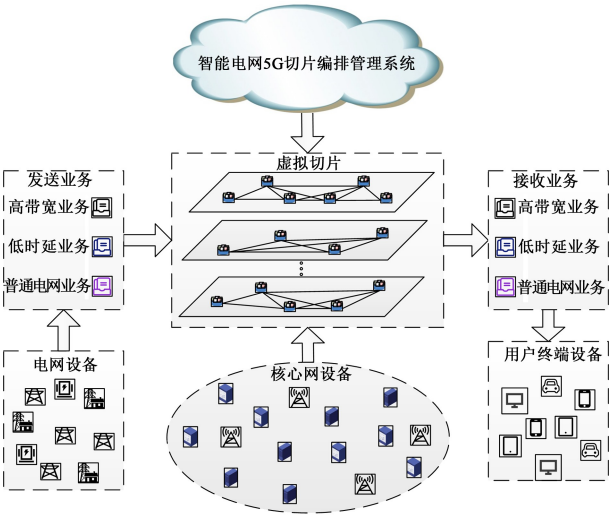


图 1　智能电网光网络切片场景

1.3　资源分配

VNF 部署后，需要执行资源分配过程。资源分配过程包括节点的计算和存储资源分配以及链路带宽资源分配。使用变量 $c_s^i(t)$，$b_s^i(t)$ 和 $w_s^l(t)$ 分别表示分配给 VNF i 的计算资源、存储资源和链路 l 的带宽资源。对于核心网中节点、VNF 和链路，为了保证接入设备业务 s 比特速率要求，VNF i 的计算资源、存储资源和链路 l 的带宽资源分配不能低于其业务资源需求 $c_{s,\mathrm{need}}^i(t)$，$b_{s,\mathrm{need}}^i(t)$ 和 $w_{s,\mathrm{need}}^l(t)$：

$$c_s^i(t)\geqslant R_s(t)c_i,\forall s\in \boldsymbol{S},o_s^i\in X_s \tag{6}$$

$$b_s^i(t)\geqslant R_s(t)b_i,\forall s\in \boldsymbol{S},o_s^i\in X_s \tag{7}$$

$$w_s^l(t)\geqslant R_s(t)/\eta_l,\forall s\in \boldsymbol{S},l\in \boldsymbol{L} \tag{8}$$

另一方面，所有业务分配资源之和不能超过节点与链路总资源，即

$$\sum_{s\in \boldsymbol{S}}\sum_{o_n^i\in \boldsymbol{O}_n} c_s^i(t)\leqslant C_n \tag{9}$$

$$\sum_{s\in \boldsymbol{S}}\sum_{o_n^i\in \boldsymbol{O}_n} b_s^i(t)\leqslant B_n \tag{10}$$

$$\sum_{s\in \boldsymbol{S}}\sum_{l\in \boldsymbol{L}} w_s^l(t)\leqslant W_l \tag{11}$$

同时，链路守恒应该得到保证，即

$$\sum_{n'\in N} w_s^{(n,n')}(t)\eta_{(n,n')}-\sum_{n'\in N} w_s^{(n^*,n)}(t)\eta_{(n,n')}\geqslant R_s,\ \forall n\in \boldsymbol{N}^{\mathrm{S}} \tag{12}$$

$$\sum_{n'\in N} w_s^{(n,n')}(t)\eta_{(n,n')}-\sum_{n'\in N} w_s^{(n^*,n)}(t)\eta_{(n,n')}\geqslant -R_s,\ \forall n\in \boldsymbol{N}^{\mathrm{D}} \tag{13}$$

$$\sum_{n'\in N} w_s^{(n,n')}(t)\eta_{(n,n')}-\sum_{n'\in N} w_s^{(n^*,n)}(t)\eta_{(n,n')}=0,\ \forall n\in \boldsymbol{N}^{\mathrm{R}} \tag{14}$$

由于优化问题同时具有离散变量环境(VNF 部署)和连续变量环境(资源分配)，因此需要进行简化。本文采用条件判断映射化简。由于 $u_{i,i'}^s$ 和 v_i^s 是二进制变量，其值仅可能等于 0 或 1，当 $u_{i,i'}^s$ 或 v_i^s 为 0 时，对应变量不存在资源分配。因此，可以使用条件语句建立两者之间的映射，即

$$u_{i,i'}^s(t)=\begin{cases}0,\text{if} & w_s^{(o_s^i,o_s^{i'})}(t)=0\\ 1,\text{if} & w_s^{(o_s^i,o_s^{i'})}(t)>0\end{cases} \tag{15}$$

$$v_i^s(t)=\begin{cases}0,\text{if} & c_s^i(t)=0\\ 1,\text{if} & c_s^i(t)>0\end{cases} \tag{16}$$

因此，VNF 部署过程可以通过资源分配映射获得，优化问题由 VNF 部署和资源分配的复合变量问题转换为资源分配的连续变量问题。同时，约束式(1)～(5)可以通过如下方式转换：

$$\sum_{i\notin \boldsymbol{N}_s}\sum_{n'\in \boldsymbol{N}} w_s^{(o_s^i,o_s^{i'})}(t)=0 \tag{17}$$

$$\sum_{n\notin \boldsymbol{N}_i} c_s^i(t)=0 \tag{18}$$

$$\sum_{n\in \boldsymbol{N}_i}\sum_{n'\in \boldsymbol{N}} w_s^{(o_s^i,o_s^{i'})}(t)=1 \tag{19}$$

$$\sum_{n\in \boldsymbol{N}_i} c_s^i(t)=c_s^{i'}(t),c_s^{i'}(t)\neq 0 \tag{20}$$

$$\sum_{n'\in \boldsymbol{N}} w_s^{(o_s^i,o_s^{i'})}(t)>0,\text{if}\quad c_s^i(t)>0 \tag{21}$$

1.4　优化目标

本文目标函数为优化电网公司收益。公司收益

由传输比特速率、成本和时延共同构成。其中，成本由 VNF 转发流量和执行计算成本、流量传输成本构成。时延由传输时延、计算和存储时延两部分构成。各部分表示如下：

$$C_{\mathrm{VNF}}(t)=\sum_{s\in\boldsymbol{S}}\sum_{n\in\boldsymbol{N}}num_u\cdot c_u+\sum_{s\in\boldsymbol{S}}\sum_{n\in\boldsymbol{N}}num_v\cdot c_v \tag{22}$$

$$C_{\mathrm{link}}(t)=\sum_{s\in\boldsymbol{S}}\sum_{l\in\boldsymbol{L}}w_s^l(t)\cdot c_l \tag{23}$$

$$\tau_{\mathrm{VNF}}(t)=\sum_{s\in\boldsymbol{S}}\sum_{n\in\boldsymbol{N}}\frac{c_{s,\mathrm{need}}^i(t)}{c_s^i(t)}+\sum_{s\in\boldsymbol{S}}\sum_{n\in\boldsymbol{N}}\frac{b_{s,\mathrm{need}}^i(t)}{b_s^i(t)} \tag{24}$$

$$\tau_{\mathrm{link}}(t)=\sum_{s\in\boldsymbol{S}}\sum_{l\in\boldsymbol{L}}\frac{w_{s,\mathrm{need}}^l(t)}{w_s^l(t)} \tag{25}$$

其中，c_u 和 c_v 表示单个 VNF 的转发流量和执行功能的成本，num_u 和 num_v 分别表示 $u_{i,i'}^s(t)$ 和 $v_i^s(t)$，即 $w_s^l(t)$ 和 $c_s^i(t)$ 不为 0 的个数。因此，目标收益可以表示为

$$R(t)=\sum_{s\in\boldsymbol{S}}R_s(t)\cdot r_s(t)-\alpha(C_{\mathrm{VNF}}(t)+C_{\mathrm{link}}(t))-\beta(\tau_{\mathrm{VNF}}(t)+\tau_{\mathrm{link}}(t)) \tag{26}$$

其中，$r_s(t)$ 表示 1 bit 业务支付报酬，α 和 β 是归一化系数。综上，优化问题可以表示为

$$\max_{c_s^i(t),b_s^i(t),w_s^l(t)} R(t),\ \text{s.t.}\ (6)-(21) \tag{27}$$

2 资源分配算法设计

2.1 状态、动作与奖励设计

本文状态定义为 VNF 剩余计算资源和存储资源、链路剩余带宽资源以及业务比特速率的集合，即 $s(t)=\{C_n^{\mathrm{r}}(t),B_n^{\mathrm{r}}(t),W_l^{\mathrm{r}}(t),R_s(t)\}$。

动作定义为每个业务分配的计算、存储和带宽资源的集合，即 $a=\{c_s^i(t),b_s^i(t),w_s^l(t)\}$。

奖励定义为目标收益与约束惩罚的差值，即 $R(t)-\mathrm{flag}\cdot r_{\mathrm{p}}$，其中 flag 是不满足的约束数量，$r_{\mathrm{p}}$ 是单条约束不满足的奖励惩罚。

2.2 基于 MADDPG 的资源分配算法

由于 DQN、深度双 Q 网络（Double Deep Q-learning Network，DDQN）和竞争构架 Q 网络（Dueling DQN）等算法都是无模型算法，Q 值以较低方差的形式估计得到。此类算法无法处理连续动作空间问题。为了解决此类问题，基于策略梯度的强化学习方法通过学习确定性或随机策略来处理具有连续动作空间的问题，策略更新通过奖励期望的梯度实现。然而，此类方法的收敛速度非常慢[15]。为了解决这个问题，深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）算法融合确定性策略和基于动作价值的 DQN 算法，可以更好地处理连续和大的状态/动作空间[15-16]。

本文的优化问题是具有连续动作空间的资源分配问题，因此 DDPG 算法成为一个很好的解决方案。然而，面对大规模和差异化的电力业务，极长的算法收敛和训练时间严重影响了网络的性能，因此需要在 DDPG 算法基础上改进。

近年来，分布式 DRL 算法成为一个研究热点。分布式架构算法能够极大减少智能体训练时间，加速算法训练和收敛。多智能体 RL 是分布式 RL 的一种，是多智能体系统与强化学习的结合。系统中每个智能体遵循强化学习的目标，并独立地与环境交互，根据环境反馈改变策略以最大化获得累计回报。由于环境状态的改变与所有智能体的联合动作相关，因此需要考虑联合动作对环境产生的影响。

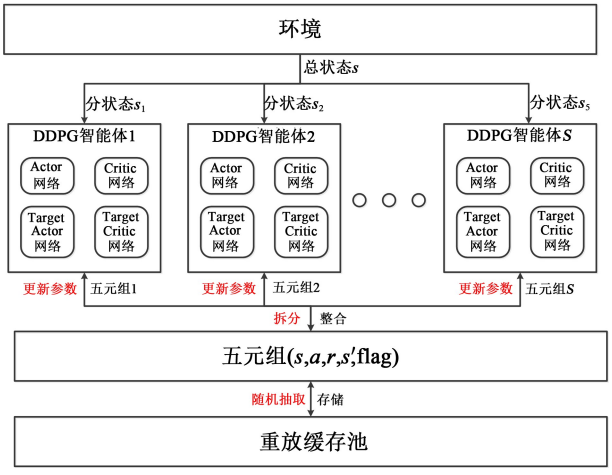


图 2 基于 MADDPG 的资源分配算法框架

在 MADDPG 中，本文将环境状态空间分为全局状态与局部状态。其中全局状态是环境的总状态，即所有节点的计算资源和存储资源，所有链路的带宽资源和所有业务的比特速率集合。在此算法中，假设每个智能体针对一种业务的资源分配，并且能够访问并调用网络所有资源，因此，局部状态为所有节点的计算资源和存储资源，所有链路的带宽资源和该业务的比特速率集合。每个智能体的动作是为该业务选择节点 VNF 和链路，并对其分配计算资源、存储资源和带宽资源。因此，联合动作是所有业务分配的计算资源、存储资源和带宽资源之和。由于每个智能体都能够分配所有网络资源，存在联

合动作的分配资源超过网络总资源、链路流量分配不守恒等问题，此时，需要通过极大地奖励惩罚机制确保业务资源分配合理。基于 MADDPG 的资源分配算法框架如图 2 所示。算法具体过程如表 1。

表 1　基于 MADDPG 的资源分配算法

算法 1 基于 MADDPG 的资源分配算法
(1) 初始化 Actor 网络和 Critic 网络参数 θ_j，φ_j　$j=1,2,\cdots S$.
(2) 初始化目标网络参数 $\theta'_j \leftarrow \theta_j$，$\varphi'_j \leftarrow \varphi_j$　$j=1,2,\cdots,S$.
(3) 初始化重放缓存池 $\boldsymbol{B}$.
(4) for $t=1$ 到最大迭代次数：
(5) for $j=1\sim S$：
(6) 每个 Actor 选取动作，并计算局部奖励
(7) end for.
(8) 计算全局奖励和下一状态 s'，并将元组存入重放缓存池中，$s=s'$.
(9) 更新过程：从重放缓存池中中采用 mini-batch 个样本。
(10) for $j=1\sim S$：
(11) 每个 agent 计算 $y_j=r_j+\gamma Q'(s'_j,a'_j \mid \theta'_j)$，计算损失函数，并根据梯度反向传播更新 Critic 网络参数 θ_j.
(12) 根据梯度反向传播更新 Actor 网络参数 φ_j
(13) for 每隔 C 次迭代：
(14) 软更新目标 Critic 网络参数 $\theta'_j \leftarrow \rho\theta_j+(1-\rho)\theta'_j$.
(15) 软更新目标 Actor 网络参数 $\varphi'_j \leftarrow \rho\varphi_j+(1-\rho)\varphi'_j$.
(16)　　end for
(17)　end for
(18) end for

3　仿真分析

为了验证算法的有效性，本文将提出的基于 MADDPG 的资源分配算法，与传统 DDPG 算法进行比较，通过对比两种算法的奖励、时延、成本，证明基于 MADDPG 的资源分配算法性能优于 DDPG 算法。另一方面，通过对比两种算法的时隙训练时间和总训练时间，证明基于 MADDPG 的资源分配算法能够极大减少算法训练时间，满足网络有效性。仿真参数如表 2 所示。

表 2　仿真主要参数

参数符号	参数含义	参数值
S	业务数量	3,6,12,24
N	网络中节点数量	10,20,40,80
L	网络中链路数量	24,48,96,192
C_n	节点最大计算资源	1～10 GHz 间的随机数
B_n	节点最大存储资源	1～10 TB 间的随机数
W_l	链路最大带宽资源	100～1 000 Mbps 间随机数
R_s	业务比特速率	90～950 Mbps 间随机数
τ	时隙间隔	0.5 ms
T	算法一次迭代包含时隙	16

基于 MADDPG 的资源分配算法和 DDPG 算法的奖励对比如图 3。从图中可以看出，基于 MADDPG 的资源分配算法奖励远高于 DDPG 算法，并且收敛速度更快。由于算法奖励极易受到约束条件影响，开始时算法并未收敛到最优解，而是收敛到次优解。随着算法不断探索，在约 400 次迭代时寻找到了最优解，获得更大的奖励，并一直保持收敛直至迭代结束。即使算法处于次优解，基于 MADDPG 的资源分配算法获得的奖励也远高于 DDPG 算法，从而证明算法的有效性。

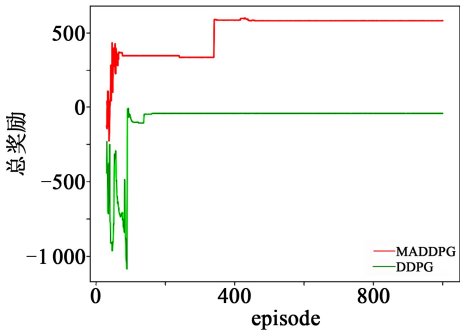


图 3　不同算法下的总奖励

图 4 是基于 MADDPG 的资源分配算法和 DDPG 算法成本的对比。图 4 显示，基于 MADDPG 的资源分配算法具有更低的成本，其物理含义为网络总资源分配更少。但是，由于基于 MADDPG 的资源分配算法奖励远高于 DDPG 算法奖励，可以推断出 DDPG 算法分配了大量无用资源或违反了过多的约束。无论是哪种情况，都可以证明基于 MADDPG 的资源分配更有效。

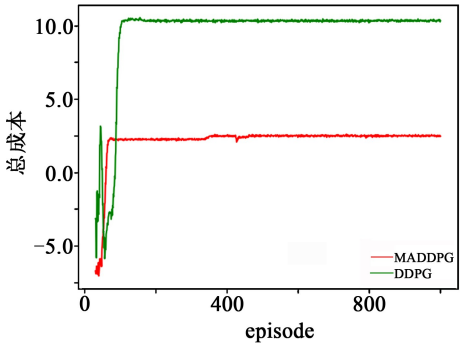


图 4　不同算法下的成本

图 5 是基于 MADDPG 的资源分配算法和 DDPG 算法时延的对比。相比成本，MADDPG 能够更有效地降低时延。更进一步，由于基于 MADDPG 的资源分配算法在 400 次迭代左右奖励有明显的上升，但是时延和成本都没有明显的变化，证明算法虽然没有增加资源使用总量，但是探索了更合理的资源分配方式，由于满足约束条件和合理的资源分配方案，使奖励明显增大。因此，此处突然上升也证明了奖励中约束惩罚项的重要性，即满足约束可以提供更加合理的资源分配，减少资源浪费。

综上所述，基于 MADDPG 的资源分配算法在

奖励、时延和成本方面具有更好的性能。同时，该算法具有更合理的业务分配方案和业务适应性，能够处理差异化电网业务。

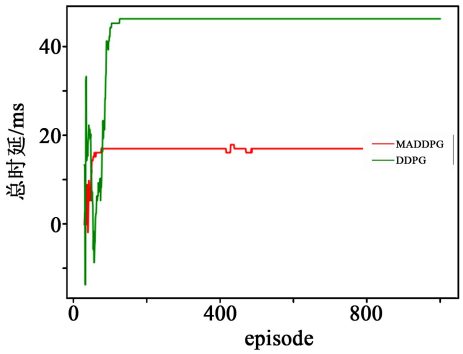


图 5 不同算法下的时延

图 6 和图 7 是针对基于 MADDPG 的资源分配算法和 DDPG 算法处理大规模业务的能力对比结果。可以看出，随着业务数量的增多，基于 MADDPG 的资源分配算法优势越发显著。当业务数量为 3 时，由于业务数量较少，DDPG 算法能够及时处理，而基于 MADDPG 的资源分配算法需要节点之间相互通信，因此基于 MADDPG 的资源分配算法训练时间长于 DDPG 算法。随着业务数量的增加，基于 MADDPG 的资源分配算法优势逐渐体现。当业务数量为 6 时，基于 MADDPG 的资源分配算法训练时间低于 DDPG 算法，并且伴随业务数量的增加差距逐渐拉大。当业务数量为 24 时，基于 MADDPG 的资源分配算法训练时间已经小于 DDPG 算法的一半。

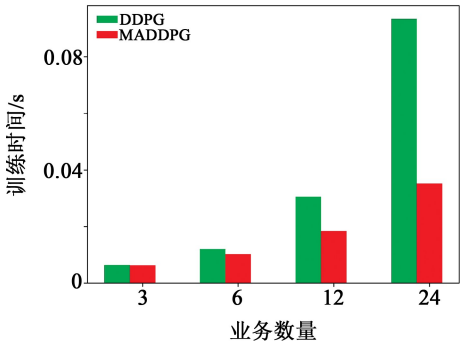


图 6 不同算法下的时隙训练时间对比

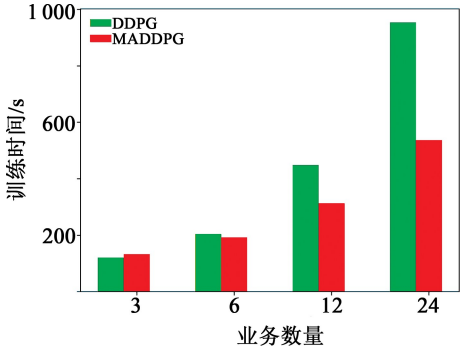


图 7 不同算法下的总训练时间对比

综上所述，基于 MADDPG 的资源分配算法在处理大规模业务和差异化业务方面都优于 DDPG 算法，并且伴随业务数量的增多，优势更为明显。未来网络中大规模和高差异化的业务将成为主流，因此本文提出的算法比 DDPG 算法更有效。

4 结论

本文从提高智能电网光网络对大规模、差异化电力业务的资源分配能力出发，提出了基于 MADDPG 的资源分配算法。本文考虑 VNF 部署和资源分配过程网络约束，建立了最大化电网公司收益为目标的优化问题。利用条件判断映射，将复合变量优化问题转换为连续变量优化问题，并提出基于 MADDPG 的资源分配算法，求解优化问题最优解。仿真结果验证了本文算法具有较好的奖励、时延和成本，并且能够大大降低大规模业务的算法训练时间，在大规模和高差异化业务为主导的未来网络中具有很好的应用前景。

参考文献：

[1] Zhang D, Han X, Deng C. Review on the research and practice of deep learning and reinforcement learning in smart grids[J]. CSEE Journal of Power and Energy Systems, 2018, 4(3): 362-370.

[2] Wang R, Aghvami A H, Friderikos V. Service-aware design policy of end-to-end network slicing for 5G use cases[J]. IEEE Trans. on Network and Service Management, 2022, 19(2): 962-975.

[3] Cui Y, Huang X, He P, et al. QoS guaranteed network slicing orchestration for internet of vehicles[J]. IEEE Internet of Things J., 2022, 9(16): 15215-15227.

[4] Li R, Zhao Z, Sun Q, et al. Deep reinforcement learning for resource management in network slicing[J]. IEEE Access, 2018, 6: 74429-74441.

[5] Albonda H D R, Pérez-Romero J. An efficient RAN slicing strategy for a heterogeneous network with eMBB and V2X services[J]. IEEE Access, 2019, 7: 44771-44782.

[6] Sun G, Gebrekidan Z T, Boateng G O, et al. Dynamic reservation and deep reinforcement learning based autonomous resource slicing for virtualized radio access networks[J]. IEEE Access, 2019, 7: 45758-45772.

[7] Sun G, Xiong K, Boateng G O, et al. Autonomous resource provisioning and resource customization for mixed traffics in virtualized radio access network[J]. IEEE Systems J., 2019,

13(3)：2454-2465.

[8] Qi C，Hua Y，Li R，et al. Deep reinforcement learning with discrete normalized advantage functions for resource management in network slicing[J]. IEEE Commun. Lett.，2019，23(8)：1337-1341.

[9] Sun J，Zhu G，Sun G，et al. A reliability-aware approach for resource efficient virtual network function deployment[J]. IEEE Access，2018，6：18238-18250.

[10] Qu L，Assi C，Shaban K，et al. A reliability-aware network service chain provisioning with delay guarantees in NFV-enabled enterprise datacenter networks[J]. IEEE Trans. on Network and Service Management，2017，14(3)：554-568.

[11] Wang D，Wang Y，Dong S，et al. On delay-aware resource control with statistical QoS provisioning by dual connectivity in heterogeneous aeronautical network[J]. IEEE Trans. on Vehicular Technology，2020，69(3)：2915-2927.

[12] Guan W，Wen X，Wang L，et al. A service-oriented deployment policy of end-to-end network slicing based on complex network theory[J]. IEEE Access，2018，6：19691-19701.

[13] Chien H T，Lin Y D，Lai C L，et al. End-to-end slicing as a service with computing and communication resource allocation for multi-tenant 5G systems[J]. IEEE Wireless Communications，2019，26(5)：104-112.

[14] Ye Q，Li J，Qu K，et al. End-to-end quality of service in 5G networks：Examining the effectiveness of a network slicing framework[J]. IEEE Vehicular Technol. Magazine，2018，13(2)：65-74.

[15] Guo J，Liu Q，Chen E. A deep reinforcement learning method for multimodal data fusion in action recognition[J]. IEEE Signal Proc. Lett.，2022，29：120-124.

[16] Yang M，Liu N，Zuo L，et al. Dynamic charging scheme problem with actor-critic reinforcement learning[J]. IEEE Internet of Things J.，2020，8(1)：370-380.

作者简介：

戚银城(1968—)，男，河南省商水县人，博士，教授，研究方向为电力通信与信息处理技术。